Soc Choice Welf (2016) 47:501–510
DOI 10.1007/s00355-016-0974-z

ORIGINAL PAPER

Pareto-optimal matching allocation mechanisms for boundedly rational agents

Sophie Bade[1,2]

Received: 21 February 2014 / Accepted: 30 May 2016 / Published online: 2 July 2016

Abstract Is the Pareto optimality of matching mechanisms robust to the introduction of boundedly rational behavior? To address this question I define a restrictive and a permissive notion of Pareto optimality and consider the large set of hierarchical exchange mechanisms which contains serial dictatorship as well as Gale's top trading cycles. Fix a housing problem with boundedly rational agents and a hierarchical exchange mechanism. Consider the set of matchings that arise with all possible assignments of agents to initial endowments in the given mechanism. I show that this set is nested between the sets of Pareto optima according to the restrictive and the permissive notion. These containment relations are generally strict, even when deviations from rationality are minimal. In a similar vein, minimal deviations from rationality suffice for the set of outcomes of Gale's top trading cycles with all possible initial endowments to differ from the set of outcomes of serial dictatorship with all possible orders of agents as dictators.

Mathematics Subject Classification C78 · D03 · D60

I would like to thank Christoph Engel, Olga Gorelkina, Martin Hellwig, Aniol Llorente-Saguer, Michael Mandler, Paola Manzini, and two anonymous referees. I would like to thank David Bielen for thorough research assistance.

✉ Sophie Bade
sophie.bade@rhul.ac.uk

1 Royal Holloway College, University of London, Egham, United Kingdom

2 Max Planck Institute for Research on Collective Goods, Bonn, Germany

1 Introduction

Boundedly rational behavior should be expected in some of the non-market environments for which economists have designed matching mechanisms. Take kidney allocation problems as an example. One difficulty with mechanisms that match donors to recipients is that doctors are reluctant to state complete preferences over kidneys. However, the same doctors do not seem to have any problem choosing the "best" kidney for a particular patient from a given set. The limited resources available to test whether a kidney is a good match might drive this apparent contradiction.[1] Alternatively, consider the allocation of elementary school slots. The choices of a family in which the mother strategically whittles down the options before the father picks a school are only rationalizable if the parents' preferences are aligned.[2] As a third example consider the choice of a medical residency program. To reduce the complexity of the choice-problem a med-school graduate might use a sequence of incomplete rankings to eliminate all but a few alternatives which she then considers in detail.[3]

Is the Pareto optimality of matching mechanisms robust to the introduction of boundedly rational behavior? To answer this question, I consider Papai's (2000) hierarchical exchange mechanisms, which comprise many theoretically and practically relevant matching mechanisms.[4] I derive two different preference-relations from choice functions. An agent lightly prefers x to y if he chooses x from *some* set that also contains y; he solidly prefers x to y if he *never* chooses y when x is also available. While the agents' light preferences imply a restrictive notion of Pareto optimality, their solid preferences imply a permissive notion of Pareto optimality.

In line with standard matching theory I find that any Pareto optimum that satisfies the restrictive notion can be obtained as the outcome of any fixed hierarchical exchange mechanism for some initial endowment and that any outcome of hierarchical exchange satisfies the permissive notion of Pareto optimality. In contrast to standard matching theory I find that the set of outcomes of hierarchical exchange is strictly nested between the two Pareto sets and that different hierarchical exchange mechanisms cover different sets of outcomes. Agents do not have to stray far from rational behavior for these two results two hold; in fact I show that minimal deviations from rational behavior suffice.

[1] These statements reflect a private conversation with Utku Unver, who was involved in the design and practical implementation of several kidney exchange mechanisms. Consider the task of choosing the "best" kidney for a patient from a set $S = \{a, b, \dots\}$ of ten kidneys. Due to financial constraints doctors may use preliminary tests to limit the set of kidneys which they examine in detail. If b is eliminated by the preliminary tests, while b turns out to be better than a according to the detailed examination, this procedure may yield the choices $a = c(S)$ and $b = c(\{a, b\})$.

[2] Xu and Zhou (2007) as well as Apesteguia and Ballester (2013) characterized choice function that can be explained via such strategic interplay of different agents.

[3] Manzini and Mariotti (2007) and Mandler (2015) characterize choice functions that arise out of such procedures.

[4] Some subsets of the class of hierarchical exchange mechanisms have been described by Abdulkadiroglu and Sönmez (1999), Svensson (1999), Ergin (2000), Ehlers et al. (2002), Ehlers and Klaus (2004), Kesten (2009), Ehlers and Klaus (2007), and Velez (2014).

2 Matchings and hierarchical exchange

Fix a set of agents, $N = \{1, \ldots n\}$, and a set H of equally many objects, called houses. A **submatching** $\sigma : N_\sigma \to H_\sigma$ is a bijection with $N_\sigma \subset N$ and $H_\sigma \subset H$; $\sigma(i)$ is agent i's match under σ. Any submatching σ is also interpreted as a set of agent-house pairs: $\{(i, h) : \sigma(i) = h\}$. If $N_\sigma \cap N_{\sigma'} = \emptyset = H_\sigma \cap H_{\sigma'}$ then $\sigma \cup \sigma' : N_\sigma \cup N_{\sigma'} \to H_\sigma \cup H_{\sigma'}$ maps i to $\sigma(i)$ if $i \in N_\sigma$ and to $\sigma'(i)$ otherwise. If $N_\mu = N$, then μ is a **matching**. Matchings are also denoted as vectors with the understanding that the ith component of μ represents $\mu(i)$. The sets of all matchings and respectively of all submatchings, that are not themselves matchings, are $\mathcal{M}$ and $\overline{\mathcal{M}}$. The submatching that matches no one, $\emptyset$, is an element of $\overline{\mathcal{M}}$. The sets of unmatched agents and houses at some $\sigma \in \overline{\mathcal{M}}$ are denoted $\overline{N}_\sigma$ and $\overline{H}_\sigma$.

I use Pycia and Unver's (2014) ingenious terminology to define Papai's (2000) **hierarchical exchange mechanisms**. For any fixed $\sigma \in \overline{\mathcal{M}}$ define an **ownership function** $o_\sigma : \overline{H}_\sigma \to \overline{N}_\sigma$, with the understanding that agent $o_\sigma(x)$ owns house x at the submatching σ. Any set of ownership functions $o = (o_\sigma)_{\sigma \in \overline{\mathcal{M}}}$ where $o_\sigma(x) = o_{\sigma'}(x)$ holds for any two submatchings $\sigma \subset \sigma'$ with $o_\sigma(x) \notin \overline{N}_{\sigma'}$ and $x \notin H_{\sigma'}$ defines a hierarchical exchange mechanism. So ownership persists in the sense that agent $i \notin N_{\sigma'}$ owns house $x \notin H_{\sigma'}$ at σ' if i owns x at a submatching σ of σ'. The outcome of any hierarchical exchange mechanism is determined through the following trading process.[5]

To begin let $\sigma_1 = \emptyset$ and $k = 1$. Round k: each house $h \in \overline{H}_{\sigma_k}$ points to its owner $o_{\sigma_k}(h)$, each agent $i \in \overline{N}_{\sigma_k}$ points to a house in $\overline{H}_{\sigma_k}$. Define σ^* as the submatching that matches each agent in some pointing cycle to the house he points to. Let $\sigma_{k+1} = \sigma_k \cup \sigma^*$. Terminate the mechanism if σ_{k+1} is a matching. If not, go on to round $k + 1$.

At the start of a hierarchical exchange mechanism, agents are asked to point to houses. Houses in turn point to their owners. At least one cycle of agents and houses forms. Any agent in such a cycle is matched with the house he points to and leaves the mechanism. When an owner of multiple houses leaves, his unmatched houses are passed on to the remaining agents according to the inheritance rule implied by the ownership functions. The remaining agents are then asked to point to the remaining houses. The procedure is repeated until each agent is matched. If $\hat{o}_\sigma(h)$ only depends on $|\ N_\sigma\ |$, the number of agents already matched under σ, then $\hat{o}$ is a **serial dictatorship**. If $\overline{o}_\emptyset(h) \neq \overline{o}_\emptyset(h')$ holds for all $h \neq h'$, then $\overline{o}$ is **Gale's top trading cycles mechanism**.

For any fixed hierarchical exchange mechanism o and any permutation $p : N \to N$ define a **permuted hierarchical exchange mechanism** $p\square o$ via $(p\square o)_\sigma(h) = p(o_{\sigma \circ p}(h))$ for all $\sigma \in \overline{\mathcal{M}}$.[6] Under $p\square o$ agent $p(i)$ takes on the role of agent i under o. If agent i is the ith dictator according to the serial dictatorship $\hat{o}$ then agent $p(i)$ is the ith dictator according to $p\square\hat{o}$. If agent 1 is endowed with houses $\{e, g, h\}$ at the

[5] The restriction to hierarchical exchange mechanisms is not costless. Pycia (2014) define a class of problems in which hierarchical exchange mechanisms are strictly Lorenz-dominated by some other strategy proof, Pareto optimal, and non-bossy mechanisms. Abdulkadiroglu et al. (2014) show that the use of ordinal mechanisms when agents have cardinal utilities may lead to welfare losses, Pycia (2014) shows that these losses can be arbitrarily large.

[6] Abusing notation let p be the restriction of the original permutation p for which $\sigma \circ p$ is well-defined.

start of some hierarchical exchange mechanism o ($o_\emptyset(e) = o_\emptyset(g) = o_\emptyset(h) = 1$), then agent $p(1)$ is endowed with these houses at the start of $p\Box o$.

3 Boundedly Rational Behavior

Fixing N and H, a **housing problem** is a profile $c := (c_i)_{i\in N}$, where $c_i : \mathcal{P}(H)\backslash\{\emptyset\} \to H$ is agent i's choice function and $c_i(S) \in S$ is agent i's choice from the set S. A choice function c_i is **rationalizable** if there exists a transitive and complete preference $\succsim_i$, such that c_i maps any $S \subset H$ to the $\succsim_i$-maximal element in S.[7] Agent i **lightly prefers** house x to house y if $x = c_i(S)$ holds for some $y \in S \subset H$; he **solidly prefers** x to y, if $x \in S \subset H$ implies $y \neq c_i(S)$. If i lightly prefers x to y I write $x P_i^\exists y$, if his preference is solid I write $x P_i^\forall y$. A matching μ' $P^\forall$**-Pareto-dominates** ($P^\exists$**-Pareto-dominates**) another matching $\mu' \neq \mu$ if $\mu'(i) \neq \mu(i)$ implies $\mu'(i) P_i^\forall \mu(i)$ ($\mu'(i) P_i^\exists \mu(i)$) for all i. A matching μ is $P^\forall$**-Pareto-optimal** ($P^\exists$**-Pareto-optimal**) if there exists no matching μ' that $P^\forall$-Pareto-dominates ($P^\exists$-Pareto-dominates) it.[8]

The mechanism o **implements** the matching $o(c)$ in a housing problem c, if $o(c)$ results when any agent at any round of the mechanism points to his choice out of all remaining houses.[9] If c is rationalizable, agent i points to his most preferred remaining house at any round. A mechanism o is said to **p-implement** a matching μ in housing problem c if $\mu = (p\Box o)(c)$ holds for some permutation p.

4 The Result

With boundedly rational behavior hierarchical exchange mechanisms are Pareto optimal in the following sense:

Theorem 1 *Fix a housing problem c and a hierarchical exchange mechanism o. Any $P^\exists$-Pareto optimum is p-implementable by o. Any matching that is p-implementable by o is $P^\forall$-Pareto optimal.*

For the proof of the first part I fix a $P^\exists$-Pareto optimum μ and show the agents can be ordered such that no agent would choose a lower ranked agent's match under μ if his own match under μ is available. To illustrate the remaining arguments assume that this ordering ranks any i above all $j > i$. Define an assignment p of agents to roles in o such

[7] Standard housing problems, profiles of linear orders $(\succsim_i)_{i\in N}$ on H, are embedded in the set of housing problems. Given that agents are represented via choice functions (not correspondences), the presence of boundedly rational behavior is the only difference between the present and the standard definition of housing problems.

[8] The notion of solid preference $P^\forall$ is identical with (or very similar to) the notions of preference that Bernheim and Rangel (2009), Mandler (2014), and Green and Hojman (2008) use to compare outcomes in terms of individual and collective welfare. Rubinstein and Salant (2012) show that this notion may not generate the relevant welfare preference.

[9] In a working paper version I show that Theorem 1 extends to more general assumptions on behavior. While my behavioral assumptions pertain to the trading process de Clippel's (2014) behavioral assumptions abstract away from the process and directly apply to the mechanism as a mapping from set of simultaneous choices to outcomes.

that agent i controls $\mu(i)$ at the submatching of μ that matches the $i-1$ highest ranked agents: $\{(1, \mu(1)), (2, \mu(2)), \ldots, (i-1, \mu(i-1))\}$. Assume for now that exactly one pointing cycle forms at each round of the mechanism $p\Box o$ at c. By the definition of p agent 1 owns house $\mu(1)$ when the mechanism starts. By the construction of the ordering agent 1 chooses $\mu(1)$ out of the set of all houses. So $\mu(1)$ and 1 form a cycle and the submatching $(\{1, \mu(1)\})$ is reached in the first round. By the definition of p agent 2 owns house $\mu(2)$ at $\{(1, \mu(1))\}$; by the construction of the ordering 2 chooses $\mu(2)$ out of all remaining houses and $\{(1, \mu(1)), (2, \mu(2))\}$ is reached in the second round. Proceeding inductively, $\mu = \{(1, \mu(1)), (2, \mu(2)), \ldots, (n, \mu(n))\}$ is reached in the nth (and last) round. The proof adapts the above arguments to the general case with any ordering and multiple cycles in one round.

Proof To prove the first part fix a $P^{\exists}$-Pareto-optimal μ. Then, I claim there exists an ordering $f : \{1, \ldots, n\} \to \{1, \ldots, n\}$ of the agents such that $\mu(f(i)) \in S$ and $j > i$ imply $c_{f(i)}(S) \neq \mu(f(j))$. So the ith agent (according to the ordering f) never chooses a match $\mu(f(j))$ of a lower ranked agent $f(j)$ with $j > i$ if $\mu(f(i))$, his own match under μ, is available. To see this suppose there was no agent i^*, who chooses $\mu(i^*)$ whenever it is available. So suppose that for each agent i there exists a set $S_i \subset H$, such that $\mu(i) \in S_i$ and $\mu(i) \neq c_i(S_i)$. Now let each agent i point to the agent who is matched with $c_i(S_i)$ under μ. The matching μ', with $\mu'(i) = c_i(S_i)$ for any agent i in some pointing cycle and $\mu(i) = \mu'(i)$ otherwise, $P^{\exists}$-Pareto dominates μ, a contradiction. So some agent i^* chooses $\mu(i^*)$ whenever it is available. Set $f(1) := i^*$. Since the restriction of μ to $N \setminus \{f(1)\}$ and $H \setminus \{\mu(f(1))\}$ is also $P^{\exists}$-Pareto-optimal, the inductive application of the above arguments implies the existence of the ordering f.

For each i define μ^i as the submatching of μ that matches the first $i-1$ agents according to the ordering f, so $\mu^i := \{(f(1), \mu(f(1))), (f(2), \mu(f(2))) \ldots (f(i-1), \mu(f(i-1)))\}$. Define p such that $(p\Box o)_{\mu^i}(\mu(f(i))) = f(i)$ for all $i \in N$.[10] So p is such that the ith agent in the ordering owns his match under μ at the submatching of μ that matches all agents who are ordered before him.

I show next that any round of $p\Box o$ at c that starts with a submatching $\sigma \subset \mu$ must end with a submatching $\sigma' \subset \mu$. Fix any $\sigma \subset \mu$. To see that house $\mu(f(i)) \in \overline{H}_\sigma$ is owned by an agent $f(j)$ with $i \geq j$, suppose some agent $f(j)$ owns a house $\mu(f(i))$ with $j \geq i$ (so $(p\Box o)_\sigma(\mu(f(i))) = f(j)$). Since $\sigma \subset \mu$ and $\mu(f(i)) \notin H_\sigma$ agent $f(i)$ is not matched at σ. Since $p\Box o$ is a hierarchical exchange mechanism and since neither house $\mu(f(i))$ nor agents $f(i)$ and $f(j)$ are matched under $\mu^i \cup \sigma$ we obtain

$$f(j) = (p\Box o)_\sigma(\mu(f(i))) = (p\Box o)_{\sigma\cup\mu^i}(\mu(f(i))) = (p\Box o)_{\mu^i}(\mu(f(i))) = f(i),$$

where the last equality follows from the definition of p. We can conclude that $(p\Box o)_\sigma(\mu(f(i))) = f(j)$ implies $i \geq j$ and any house $\mu(f(i))$ points to an agent

[10] To see that p is well-defined note that p declares agent $f(i)$ to be the owner of $\mu(f(i))$ at μ^i. Since any $\mu(f(j))$ with $j < i$ is matched under μ^i the role of owner of $\mu(f(i))$ at μ^i differs from the role of owner of $\mu(f(j))$ at $\mu^j \subset \mu^i$ for any $j < i$. So p specifies a role for agent $f(i)$ that differs from the all roles to which p assigns the agents $f(j)$ with $j < i$. Since there are equally many roles as there are agents, p is a well-defined bijection.

$f(j)$ with $i \geq j$. Since $c_{f(j)}(\overline{H}_\sigma) = \mu(f(i))$ implies $i \leq j$ any unmatched agent $f(j)$ points to a house $\mu(f(i))$ with $i \leq j$. Consequently any cycle at σ involves just one agent $f(j)$ and his match $\mu(f(j))$ and any round starting with a submatching $\sigma \subset \mu$ ends with a submatching $\sigma' \subset \mu$. Since the trading process starts with $\emptyset \subset \mu$ and since it must end with a matching, we obtain $(p\Box o)(c) = \mu$ and thereby the first part of Theorem 1.

To see the second part of Theorem 1 fix any $\mu = (p\Box o)(c)$. Assume w.l.o.g. that agents $\{1, \ldots, j\}$ are matched in the first round of the mechanism. So for each $i \leq j$ $\mu(i)$ is $P_i^\forall$-optimal in H. By the same argument, house $\mu(i)$ is $P_i^\forall$-optimal in $H \setminus \{\mu(1), \ldots, \mu(j)\}$ if agent i is matched in the second round. Proceeding inductively, we see that μ is $P^\forall$-Pareto-optimal. □

To see that the set of matchings implementable through hierarchical exchange is generally strictly nested between the sets of $P^\exists$- and $P^\forall$-Pareto optima consider the following two examples. Example 1 shows that some $P^\exists$-Pareto-inferior matchings are p-implementable by any hierarchical exchange mechanism. Example 2 shows that not every $P^\forall$-Pareto-optimal matching is p-implementable. For both examples let $H = \{x, y, z, w\}$, $N = \{1, 2, 3, 4\}$ and let $x \succ_i^* y \succ_i^* z \succ_i^* w$ rationalize the choice function c_i^*. Arbitrarily fix all choices that are not explicitly mentioned.

Example 1 Define c^α such that $c_1^\alpha := c_1^*$, $c_2^\alpha := c_2^*$, $c_3^\alpha(S) := y$ if $y \in S$, $c_3^\alpha(\{x, z, w\}) := w$, $c_3^\alpha(\{z, w\}) := z$, $c_4^\alpha(S) := x$ if $x \in S$, $c_4^\alpha(\{y, z, w\}) := z$, and $c_4^\alpha(\{z, w\}) := w$. The matching $\mu^\alpha := (x, y, z, w)$ is not $P^\exists$-Pareto-optimal in c^α since $c_3^\alpha(\{x, z, w\}) = w$ and $c_4^\alpha(\{y, z, w\}) = z$ imply $w P_3^\exists z$ and $z P_4^\exists w$. Fix any hierarchical exchange mechanism o and define p such that 1 initially owns x and 2 owns y at the submatching $\{(1, x)\}$, formally $(p\Box o)_\emptyset(x) := 1$ and $(p\Box o)_{\{(1,x)\}}(y) := 2$. If there are exactly two owners at $\{(1, x)\}$ under $p\Box o$, 3 is the other owner, if there are three owners, let $(p\Box o)_{\{(1,x)\}}(w) = 4$. In the first round of the mechanism agents 1, 2, and 4 point to x while 3 points to $y = c_3^\alpha(H)$. Each house h points to its owner $(p\Box o)_\emptyset(h)$, so house x points to 1. Moreover, y cannot point to 3, since $(p\Box o)_{\{(1,x)\}}(y) = 2$ implies $(p\Box o)_\emptyset(y) \neq 3$. Exactly one cycle forms, and the submatching $\{(1, x)\}$ is reached. At $\{(1, x)\}$ agents 2 and 3 point to $y = c_2^\alpha(\{y, z, w\}) = c_3^\alpha(\{y, z, w\})$. Given that agent 2 owns house y at $\{(1, x)\}$, agent 2 and y form a cycle. This is the only cycle: if 4 owns a house at $\{(1, x)\}$ he points to $c_4^\alpha(\{y, z, w\}) = z$ which is owned by 3 at $\{(1, x)\}$. Only 3, 4, z and w are left in the next round. Since $c_3^\alpha(\{z, w\}) = z$ and $c_4^\alpha(\{z, w\}) = w$ the desired matching obtains: $(p\Box o)(c^\alpha) = \mu^\alpha$.

Example 2 Define c^β such that $c_i^\beta := c_i^*$ for $i \neq 3$, $c_3^\beta(S) := y$ if $y \in S$, $c_3^\beta(\{x, z, w\}) := z$, and $c_3^\beta(\{z, w\}) := w$. The matching $\mu^\beta := (x, y, z, w)$ is $P^\forall$-Pareto optimal in c^β. Suppose $o(c^\beta) = \mu^\beta$ held for some hierarchical exchange mechanism o. Since $c_i^\beta(H)$ equals x for agents $i = 1, 2$, and 4 while it equals y for agent $i = 3$ and since the first round must produce a submatching $\sigma \subset \mu^\beta$ only agent 1 and house x are matched in that first round. By the same logic, only agent 2 and house y are matched in the second round. In the third round, agents 3 and 4 point to $c_3^\beta(\{z, w\}) = w$ and $c_4^\beta(\{z, w\}) = z$, contradicting $o(c^\beta) = \mu^\beta$.

The next example shows that serial dictatorship and Gale's top trading cycles may p-implement different sets of matchings with boundedly rational agents. The example sheds some light on possible extensions of the growing literature on the equivalence between random serial dictatorship and other random matching mechanisms. According to random serial dictatorship the order of all agents as dictators is drawn from a uniform distribution over all such orders. Abdulkadiroglu and Sönmez (1998) and Knuth (1996) independently found that random serial dictatorship is identical to the "core from random endowments" which starts Gale's top trading cycles from an endowment that has been randomly drawn from a uniform distribution over all possible endowments.[11] Example 3 shows that the supports of the two random matching mechanisms differ with boundedly rational behavior.

Example 3 Let $H = \{x, y, z\}$, $N = \{1, 2, 3\}$ and define c^γ such that $c_1^\gamma(\{x, y, z\}) := x$, $c_1^\gamma(\{x, z\}) := z$, $c_2^\gamma(\{x, y, z\}) := y$, and $c_2^\gamma(\{y, z\}) := z$. Let $x \succ_3^\gamma y \succ_3^\gamma z$ rationalize c_3^γ. Gale's top trading cycles with the initial endowment $\mu^\gamma := (x, y, z)$ implements μ^γ in c^γ given that agents 1 and 2 choose $\mu^\gamma(1)$ and $\mu^\gamma(2)$ out of the grand set. For μ^γ to be p-implementable via serial dictatorship either 1 or 2 has to be the first dictator. But if either 1 or 2 is the first dictator, neither one of the two remaining agents would pick $\mu^\gamma(i)$ as the second dictator.

The choice behavior assumed in Examples 1, 2, and 3 is not wildly irrational. Quite to the contrary the behavior in each of these examples only minimally deviates from rationality. To make this statement precise requires a formal way of measuring the degree of irrationality. However, different theories in the literature use different measures of irrationality. Behavior that is sequentially rationalizable following Manzini and Mariotti (2007) is minimally irrational if two rationales suffice to explain it. Behavior that can be explained as choices via checklist following Mandler (2015) is minimally irrational if the checklist has length two. The minimal game tree that may explain boundedly rational behavior following Xu and Zhou (2007) has two agents and two nodes. Kalai (2002), Ambrus and Rozen (2015), Apesteguia and Ballester (2014) and Manzini and Mariotti (2012) define yet further measures of irrationality.

All these theories agree that c_i has to violate WARP at least once to qualify as boundedly rational. To judge whether c_i is minimally irrational according to the theories mentioned above we need to know $c_i(S)$ for a variety choice sets S. Consider a choice set with three elements. Appropriately renaming of the choice set as $X = \{x, y, z\}$, $c_i(\{x, y, z\}) = x$ and $c_i(\{x, y\}) = y$ must hold for c_i to violate WARP. For c_i to be minimally irrational some theories then require the choice $c_i(\{y, z\}) = z$; others do not.[12] What stands out about Examples 1, 2, and 3 is that all choice functions in these examples are either rationalizable or they violate WARP exactly once. The omission of some (arbitrarily fixed) choices turns out to be more than a notational convenience. These omitted choices were not used to establish any of the points made

[11] This result has been extended to larger sets of mechanisms by Carroll (2014), Pathak and Sethuraman (2011), Pycia and Liu (2013) and Bade (2014).

[12] For c_i to be rationalizable by two sequential rationales following Manzini and Mariotti (2007) $c_i(\{y, z\}) = z$ must hold. However, in the framework of Kalai (2002), two rationales suffice to rationalize c_i, whether we let $c_i(\{y, z\}) = z$ or $c_i(\{y, z\}) = y$.

in the examples and we may fix them to fit any desired notion of minimal irrationality. In sum we obtain that, no matter how little irrationality we permit in housing problems and no matter which theory we use to measure the degree of irrationality, some $P^{\forall}$−Pareto-optimal matchings are not implementable by any hierarchical exchange mechanism and some $P^{\exists}$-Pareto dominated matchings are p-implementable by any hierarchical exchange mechanism. Finally, serial dictatorship and Gale's top trading cycles p-implement different sets of matchings - even if we allow only minimal deviations from rationalizability.

5 Conclusion

Hierarchical exchange mechanisms can be viewed as a version of *free trade* in matching environments, where indivisible goods have to be matched to agents without recourse to prices. At any moment in the mechanism, each house is owned by someone, in the sense that the owner can freely appropriate or exchange the house. Hierarchical exchange mechanisms allow for a broad and fine spectrum of initial endowments, ranging from maximal to minimal inequality (from serial dictatorship to Gale's top trading cycles).[13] Since the results presented here hold for *all* hierarchical exchange mechanisms, they automatically hold for any subset thereof. The results apply in particular if we adopt a more restrictive notion of *free trade* for matching environments such as Abdulkadiroglu and Sönmez' Abdulkadiroglu and Sönmez (1999) top trading cycles mechanisms or Gale's top trading cycles.

Identifying hierarchical exchange mechanisms with free trade the main results of the paper can be interpreted as versions of the first and second fundamental theorem of welfare economics for the case of boundedly rational behavior.[14] The second part of Theorem 1 corresponds to a First Welfare Theorem for solid preferences: any matching that arises out of free trade is $P^{\forall}$-Pareto optimal. The first part corresponds to a Second Welfare Theorem for light preferences:[15] any $P^{\exists}$-Pareto optimum can be

[13] The analogy has its limits. Owners are, for example, neither allowed to destroy their houses nor to determine the heirs of their houses as they leave the mechanism.

[14] All fundamental theorems of welfare economics with boundedly rational agents that I am aware of concern market environments with divisible goods. Bernheim and Rangel (2009) prove a First Welfare Theorem for markets that are standard except for the assumption that the agents' behavior need not be rationalizable. Their notion of Pareto optimality relies on a notion of preferences that is very similar to the solid preferences defined here. This result aligns with the first inclusion relation of Theorem 1. Interestingly, Mandler (2014) proves a version of the Second Welfare Theorem that also defines Pareto optimality with respect to $P^{\forall}$-preferences. This discrepancy is explained by Mandler's (2014) assumption that agents act fully rationally according to their solid preferences. In Mandler (2014) the choice functions only serve to construct these preferences, individuals are always willing to select any preference-maximal element of a choice set. I, in contrast, not only use the choice functions to construct the $P^{\forall}$-preferences; I also impose that for any agent's choice in a mechanism there needs to be some set that is consistent with the underlying facts, such that the agent's choice can be construed as a choice from this set. The same comments apply to the comparison between my results and the welfare theorems in Fon and Otani (1979). However there is an additional difference as Fon and Otani (1979) assumes intransitive and incomplete preferences. The assumption of such preferences rules out many irregularities that are permissible in the present framework.

[15] Due to the finiteness of matching problems this Second Welfare Theorem does without local nonsatiation or convex upper contour sets.

p-implemented by any hierarchical exchange mechanism. Examples 1 and 2 show that the respective stronger versions of the two fundamental theorems do not hold: Some matchings that arise out of free trade are not $P^{\exists}$-Pareto optimal and some $P^{\forall}$-Pareto optima cannot be achieved through free trade.

As a further step one could explicitly model the reasons for particular forms of bounded rationality and/or decision procedures. One could, for example, assume that patients do have (linear) preferences over kidneys, but that it is costly to learn these preferences. In this case the observed bounded rationality can be derived from a fully rational (but unobserved) preference. An allocation mechanism would then interact with some form of strategic information acquisition. Bade (2015) shows that serial dictatorship is the only ex ante Pareto optimal, non-bossy and strategy proof mechanism in a matching environment with endogenous information acquisition. Similarly, one could explicitly model the interaction between family members when selecting a mechanism for school choice.

References

Abdulkadiroglu A, Sönmez T (1998) Random serial dictatorship and the core from random endowments in house allocation problems. Econometrica 66:689–701

Abdulkadiroglu A, Sönmez T (1999) House allocation with existing tenants. J Econ Theory 88:233–260

Abdulkadiroglu A, Che Y-K, Yasuda Y (2014) Expanding choice in school choice. Am Econ J Microecon 7:1–42

Apesteguia J, Ballester M (2013) Choice by sequential procedures. Games Econ Behav 77:9099

Apesteguia J, Ballester M (2014) A measure of rationality and welfare. J Political Econ (**f**orthcoming)

Ambrus A, Rozen K (2015) Rationalizing choice with multi-self models. Econ J 125:1136–1156

Bade S (2015) Serial dictatorship: the unique optimal allocation rule when information is endogenous. Theor Econ 10:385–410

Bade S (2014) Random serial dictatorship: the one and only. Mimeo, Royal Holloway

Bernheim D, Rangel A (2009) Beyond revealed preference: choice theoretic foundations for behavioral welfare economics. Quart J Econ 124:51–104

Carroll G (2014) A general equivalence theorem for allocation of indivisible objects. J Math Econ 51:163–177

de Clippel G (2014) Behavioral Implementation. American Economic Review 104:2975–3002

Ehlers L, Klaus B (2004) Resource-monotonicity for house allocation problems. Int J Game Theory 32:545–560

Ehlers L, Klaus B (2007) Consistent house allocation. Econ Theory 30:561–574

Ehlers L, Klaus B, Papai S (2002) Strategy-proofness and population-monotonicity for house allocation problems. J Math Econ 38:329–339

Ergin H (2000) Consistency in house allocation problems. J Math Econ 34:77–97

Fon V, Otani Y (1979) Classical welfare theorems with non-transitive and non-complete preferences. J Econ Theory 20:409–418

Green J, Hojman D (2008) Choice, rationality and welfare measurement. Mimeo Harvard University, Cambridge

Kalai G, Rubinstein A, Spiegler R (2002) Rationalizing choice functions by multiple rationales. Econometrica 70:2481–2488

Kesten O (2009) Coalitional strategy-proofness and resource monotonicity for house allocation problems. Int J Game Theory 38:17–22
Knuth D (1996) An exact analysis of stable allocation. J Algorithms 20:431–442
Mandler M (2014) Indecisiveness in behavioral welfare economics. J Econ Behav Organ 97:219–235
Mandler M (2015) Rational agents are the quickest. J Econ Theory 155:206–233
Manzini P, Mariotti M (2007) Sequentially rationalizable choice. American Econ Rev 97:1824–1839
Manzini P, Mariotti M (2012) Choice by lexicographic semiorders. Theor Econ 7:1–23
Papai S (2000) Strategyproof assignment by hierarchical exchange. Econometrica 68:1403–1433
Pathak P, Sethuraman J (2011) Lotteries in student assignment: an equivalence result. Theor Econ 6:1–17
Pycia M, Liu Q (2013) Ordinal efficiency, fairness, and incentives in large markets. Mimeo, UCLA, Los Angeles
Pycia M, Unver U (2014) Incentive compatible allocation and exchange of discrete resources. Mimeo, UCLA, Los Angeles
Pycia M (2014)The cost of ordinality. Mimeo, UCLA, Los Angeles
Rubinstein A, Salant Y (2012) Eliciting welfare preferences from behavioral datasets. Rev Econ Stud 79:375–387
Svensson LG (1999) Strategy-proof allocation of indivisible goods. Soc Choice Welf 16:557–567
Velez R (2014) Consistent strategy-proof assignment by hierarchical exchange. Econ Theory 56:125–156
Xu Y, Zhou L (2007) Rationalizability of choice functions by game trees. J Econ Theory 134:548–556